Using Online Bioinformatics Tools and Databases in the Undergraduate Biology Curriculum

American Society for Microbiology
Undergraduate Education Conference

May 16, 2003

NCBI Sequence Databases and Entrez

There are a wide variety of deferent kinds of DNA and protein sequences found in the NCBI databases. The following exercise uses the Entrez system to examine several of the different types of sequence records available for the malaria parasite *Plasmodium falciparum* and precomputed analyses of these sequences.

Nucleotide sequences

Starting from the NCBI homepage type in the term *Plasmodium falciparum* and click go. How many records did you retrieve? Not all of these are actually sequences from *P. falciparu*m. This is because the search defaults to search [All Fields] of the record. To force Entrez to only retrieve *P. falciparum* records, click on the Limits tab and select [Organism] as the limited field from the pull-down list. Now click "Go" again with this limit in place. How many *P. falciparum* records are there?

You can use some of the other limits to sort these records into various categories. Use the molecule pull down list to select mRNA as the molecule type and click "Go". How many mRNA sequences are there? The majority of these are first pass single read sequences called expressed sequence tags (ESTs), bulk records with little of no annotation. Search for accession T02580 to see an example. Can you tell from this record what gene product this is? Click on the History tab and link back to the search with all *P. falciparum* mRNAs. Return to the Limits, check the "Exclude ESTs" checkbox, and run the search again. How many sequences are left? In the case of *P. falciparum* these will all be traditional GenBank records (well characterized and accurate sequences.) Retrieve accession number M93720 to see an example of a traditional GenBank record. What gene product is this?

Use the "History" tab to return to the search containing all *P. falciparum* records. Now use the limits tab to select genomic DNA/RNA as the molecule type and run the search. How many records are there? Many of these are another type of bulk sequence; genome survey sequences (GSSs). These are first-pass single-read genomic sequences that are generated in large quantities as a preliminary survey of clone libraries to be used in a genome sequencing project. Eliminate the GSS records using the exclude GSS checkbox on "Limits". The remaining records are of three kinds: traditional GenBank records that are the genomic equivalent of the traditional mRNA record we saw earlier; draft and finished sequences of the chromosomes for the *P. falciparum* genome; and assemblies of these into the *P. falciparum* chromosomes. Retrieve accession AL844503. This is the assembly of *P. falciparum* chromosome 4. How large is this sequence? Notice at the bottom of the record there is a CONTIG statement that describes how this sequence can be assembled from smaller units that exist in GenBank. Retrieve the first one of these units by clicking on the linked identifier (AL034557.8). This is a finished sequence from this genome project. This is the eighth version of this record. Link to a previous version of the record through the linked identifier (gi:5731897) in the COMMENT field of the record. This record is a draft genomic sequence, another type bulk GenBank record also called a high throughput genome record (HTG). The HTG division of GenBank was established to allow for rapid access to genome data even in preliminary stages. The rapid availability of draft sequence across the whole genome can provide useful information and was very important in producing the draft human genome two years ago.

In an effort to provide consistent and up to date annotation of genomes NCBI has its own versions of these *P. falciparum* chromosomes. These are one of the products of the NCBI Reference Sequence (RefSeq) project. To find these go back to the P. falciparum genomic DNA search click on the limits tab. Choose RefSeq from the "Only from" pull-down list and run the search. There are 14 chromosomes plus the mitochondrial genome. At this point the information on these is the same as that on the original GenBank records. However NCBI will maintain the annotation on these to keep them current. The RefSeq collection includes a number of different kinds or protein and nucleic acid sequences. RefSeq are easily recognized by their characteristic identifiers (accession numbers) that always have a two letter prefix followed by an underscore (e.g. NC_). A complete list of RefSeq accession numbers is given in the table at the end of this exercise.

Protein Sequences

Perform a search in the Entrez protein database to retrieve all *Plasmodium falciparum* proteins. As before, use the Limits to restrict this to the [Organism] field. Click on the Preview / Index tab and add the term multidrug resistance to your search by typing in the "Add terms to query" box. Use the pull-down to restrict to the [Title] field to get the most precise retrieval. Add this to the search by clicking on the "AND" button and then "Go". How many proteins do you retrieve? There are actually only two different proteins represented. Many of these entries are redundant sequences that have been deposited by multiple submitters or have been imported from outside protein-only sequence databases such as the Protein Information Resource (PIR) or Swiss-Prot. To get a non-redundant set of proteins use the Limits again to choose only RefSeqs. You should now have two records.

Conserved Domains

Conserved domains in proteins are recognizable sequence signatures based on conserved residues across large evolutionary distances. These conserved regions often correspond to conserved structure and function in proteins. The presence of a particular conserved domain in an unknown protein may give important clues to its function. At NCBI, conserved domains are identified by comparing the protein sequence to substitution frequency matrices from multiple sequence alignments of proteins containing these domains using a tool called Reverse PSI-BLAST or RPS_BLAST. The matrices used by RPS_BLAST are derived from several different databases: PFAM and SMART, two popular outside databases; and the NCBI Clusters of Orthologous Groups of proteins (COGs) and Conserved Domain database,

The *P. falciparum* RefSeqs retrieved above are ATP dependent transmembrane transporters (ATP binding cassette superfamily). You can easily show the presence of conserved domains identifying this protein superfamily by following the "Domains" link for NP_703574. Once you get to the display, click on the "Show "details button. This output shows the locations of conserved domains in the protein. s.. The different domain databases have different definitions for there domains; therefore the domains overlap in extent and match the sequence with different scores.

Link to one of the ABC_ATPase domains by clicking on the graphic. This is an NCBI curated domain. The default display shows a multiple sequence alignment of 10 of the most diverse proteins containing this domain. You can highlight various features of the domain using the "feature" pull-down list. Included in the alignment is a protein sequence from an experimentally determined protein structure. In this case the structure is of the ATPase subunit (malK) of the maltose transport system of *Thermococcus litoralis*. This illustrates the point that in many cases single domains from multidomain eukaryotic proteins are separate proteins in bacteria. You can display this structure as a model for the domain in Cn3D by clicking on the "View 3D structure" button. The structure is displayed along with the multiple sequence alignment and a structural alignment of other protein structures with this domain. Use the annotations panel to highlight the residues involved in the ATP binding site. Which of these conserved residues appears to be involved in coordinating the Mg^{2+} ion bound to ATP?

Related Sequences

Protein Neighbors

The related sequences link provides a simple list of all proteins in the database that are similar to the protein of interest. These are often called protein neighbors and are ranked in order from most similar to least similar. The ranking is by BLAST score. This view can be combined with Entrez searches through the History feature to get precise retrieval. However, there is no way to view the alignments or to see exactly how the proteins are related. A more advanced view called BLink that does allow access to the alignments is covered in the next section.

Return to the protein entry NP_703574. As we saw previously, this protein has a domain that is similar to the ATP binding component of the maltose transport system and other permeases of bacteria. You can use the related sequences link to find these proteins**.** Click on related sequences link to NP_703574. How many records do you retrieve? It would be very difficult to sort through these to find bacterial proteins. Notice that the first several of these are redundant (identical) sequences. The History feature of Entrez allows you to combine this related sequences search with additional terms to retrieve bacterial homologs. Click on the History link. The related sequences search will be listed as "Protein Neighbors." Take the number given for this search and combine it with an organism restricted search for *Salmonella typhimurium LT2*. For example:

#24 AND Salmonella typhimurium LT2[Organism]

Use the limits to only get RefSeqs. How many records do you retrieve? Examine the titles of these. What are the functions of these proteins? Find the ATP-binding component of the maltose transporter in this list. Follow the link to the Genome from this record. Then link through the hot linked identifier to the graphical display in Entrez genomes. Search the *S. typhimurium* genome for malK using the search box. This display shows the malK open reading frame in the context of the entire mal operon (malG – malM).

Blink

A more sophisticated view of related sequences is the BLink (BLAST link). This display provides a nonredundant list of related proteins, provides access to several different protein subsets, taxonomic information, and the alignments themselves. One drawback is that it is limited to the top 200 alignments.

Follow the Blink link from NP_703574. Notice that the list of proteins is nonredundant.. Click the "Best hits" button to reduce the redundancy even more. Now only the best hit from each species in the alignment is shown. No bacterial proteins show up in this list, however, because they are not in the top 200 high-scoring alignments for the entire database. Change the "Keep only" pull-down list to "Complete genomes" and click "Select" to see some bacterial proteins. You still won't find malK and other single domain ATP-binding components because they are not the top 200. There are bacterial proteins tat contain both the tranmembrne domain and the ATP binding domain in one protein. Click on the linked BLAST score for the highest-scoring hit for *Salmonella* to see the alignment of this with the *P. falciparum* protein. Why are there two alignments?

RefSeq accession numbers

Accession	**Molecule**	**Method**	**Note**	
NC_123456	Genomic	Curation	Complete genomic molecules including genomes, chromosomes, organelles, plasmids.	
NG_123456	Genomic	Curation	Incomplete genomic region; primarily supplied for *Homo sapiens* and *Mus musculus* to support the NCBI Genome Annotation pipeline.	
NM_123456	mRNA	Curation		
NR_123456	RNA	Curation	Non-coding transcripts including structural RNAs, transcribed pseudogenes, and others	
NP_123456	Protein	Curation		
NT_123456	Genomic	Automated	Intermediate genomic assemblies of BAC sequence data	
NW_123456	Genomic	Automated	Intermediate genomic assemblies of Whole Genome Shotgun sequence data	
XM_123456	mRNA	Automated	*Homo sapiens* model mRNA provided by the Genome Annotation process; sequence corresponds to the genomic contig.	
XR_123456	RNA	Automated	*Homo sapiens* model non-coding transcripts provided by the Genome Annotation process; sequence corresponds to the genomic contig.	
XP_123456	Protein	Automated	*Homo sapiens* model proteins provided by the Genome Annotation process; sequence corresponds to the genomic contig.	
NZ_ABCD12345678	Genomic	Automated	An ordered collection of whole genome shotgun sequence data, for incomplete bacterial genomes.Accessions are not tracked between releases.The first four characters following the underscore (e.g. 'ABCD') identifies a genome project.	
ZP_12345678	Genomic	Automated	Proteins, annotated on NZ_ accessions.	

@ Method:
Curated: indicates the process flow includes expert review for some of the records; analysis may be provided either by NCBI staff or collaborators.

Automated: indicates records that are not individually reviewed; updates are released in bulk for a genome.

Jurassic Park: Dinosaur DNA in GenBank?

In this exercise, students will use pre-computed related sequences in Entrez and nucleotide-nucleotide BLAST to identify the true source of putative dinosaur DNA.

There is one GenBank record that is labeled as dinosaur DNA. It appears to be bacterial contamination however. You can demonstrate this by using the related sequences link in Entrez.

1. Type accession number U41319 into the search box on the NCBI homepage and click go. According to the title of this record this is putative dinosaur DNA. Notice the comment on the record indicating that it shares no significant homology with any sequence in the database. Follow the related sequences link in the links pull-down. Based on these related sequences can you identify the source of the DNA?
2. Use the BLAST 2 Sequences utility linked to the NCBI BLAST page to compare U41319 with the next record in the list. This is a region of a bacterial genome. You can simply enter the two accession numbers in the BLAST 2 Sequences form. You will need to uncheck the box next to the word “Filter”. This will prevent masking of repetitive regions in the sequence. Click “Align”. What is the percent identity between these two sequences in the aligned region? Do you think this bacterium is the likely source of this contamination?

The above example illustrates that the annotations provided on GenBank records are those provided by the submitters may not be up to date.

Michael Crichton's fantasy about cloning dinosaurs, Jurassic Park, contains a putative dinosaur DNA sequence.

1. Use nucleotide-nucleotide BLAST against the default nucleotide database, nr, to identify the real source of the following sequence.

ftp://ftp.ncbi.nih.gov/pub/cooper/ASMMay2003/jurassic.txt

Select, copy and paste it into the BLAST form window.

2. NCBI scientist Mark Boguski noticed this obvious "contaminant" and supplied Crichton with a better sequence, shown below, for the sequel, The Lost World. Identify the most likely source of this sequence using nucleotide-nucleotide BLAST. Is this a better choice for a dinosaur DNA sequence? Mark imbedded his name in the sequence he provided. To see Mark's name use the translating BLAST (blastx) page with the sequence below. (Look for MARK WAS HERE NIH).

ftp://ftp.ncbi.nih.gov/pub/cooper/ASMMay2003/LostWorld.txt

Although the above two examples are artifacts, there are a number of DNA sequences in GenBank for extinct organisms. To see these follow the link from the NCBI homepage to the Taxonomy Browser. Follow the link on the left blue sidebar to extinct organisms.

Viewing and Aligning Protein Structures

Goals

In this exercise, students will compare a co-factor binding site in an enzyme from a thermophilic bacterium to that of a homolog in *E. coli*, and then by comparing the two structures attempt to find changes in the protein that allow it to be stable at high temperatures.

Introduction

This exercise will focus on an enzyme named DNA photolyase. DNA photolyase, as its name implies, is an enzyme that uses light (*photo*) to cleave (*lyase*) DNA molecules. The particular DNAs that this enzyme cleaves are DNAs that have been damaged by UV radiation so that they contain pyrimidine dimers, in which two pyrimidine bases have become bonded together. Such damage prevents the DNA from being properly used by the cell. DNA photolyase repairs such damage by using energy accepted by a photon to cleave the bond connecting the two pyrimidine bases.

Finding the Structure of DNA Photolyase from a Thermophile

First, we will use the Entrez system to find the structure record for DNA Photolyase from the thermophilic bacterium, *Thermus thermophilus*. This bacterium is often found in hot springs and prefers temperatures between 50°C and 85°C.

1. From the NCBI home page, click on the Structure link on the top tool bar. Next, enter the following query into the search box and click Go:

```
dna photolyase AND thermus thermophilus[organism]
```

2. Note that two structures are found. One is labeled as a complex with thymine, and one is not. Click on the accession of the one that is not the complex, 1IQR.
3. From the Structure Summary record, we see a variety of information about the record, such as the fact that it consists of one polypeptide chain labeled A. When was this structure submitted? About how many amino acids does chain A contain?

Viewing the Photolyase Structure and a Co-factor Binding Site

To view the structure of the DNA photolyase from *T. thermophilus*, click the View 3D Structure button. This launches the Cn3D application. Cn3D opens with two windows: one showing the structure and one showing the sequence. Also by default, the structure and sequence are colored by secondary structure, with alpha helices being green, beta strands being tan, and loop/coil regions being blue. You can manipulate the structure as follows:

Rotate: left click and drag
Zoom: hold <Ctrl> key, then left click and drag
Translate: hold <Shift> key, then left click and drag

To do its work, DNA photolyase requires two co-factors: one to accept the photon, and one to catalyze the breaking of the dipyrimidine bond. The latter co-factor is usually the flavin-adenine dinucleotide, FAD, and this co-factor is found in this structure. We will now investigate this binding site in some detail.

1. Using the controls above, manipulate the structure until you can clearly see the bound FAD co-factor. It will appear as “ball and sticks”.
2. Double click on any atom of the FAD to highlight the co-factor. It will turn yellow when you have succeeded.
3. From the menus in the structure window, choose Show/Hide / Select by Distance / Residues only, and then enter a radius of 3.0 angstroms in the box. Click OK. All residues within 3.0 angstroms of the FAD are now highlighted in yellow.
4. From the menus, choose Style / Edit Global Style. Uncheck the boxes next to Helix Objects and Strand Objects to turn these off. Check the box next to Protein sidechains to turn them on, and select “Ball and stick” from the pull down menu to the right. Next click on the Labels tab, and set the spacing to 1 under Protein Backbone. Click Done.
5. From the menus, choose Show/Hide / Show Selected Residues to show only the highlighted residues contacting the FAD co-factor. From the menus, choose Style / Coloring Shortcuts / Molecule. Now click anywhere in the sequence window to remove the highlighting. The residues comprising the binding site should now be purple, both in the structure and in the sequence windows. Make a note of these residues. Quit Cn3D when you are finished.

Finding the Homolog of DNA Photolyase in *E. coli*

We will now use a protein BLAST search to find the protein structure record from *E. coli* that has the highest sequence similarity to the *T. thermophilus* DNA photolyase.

1. Go back to the NCBI home page, and click the BLAST link on the top tool bar. On this page, click Standard Protein-protein BLAST.
2. Enter 1IQRA, the accession of chain A in 1IQR, in the Search box. Next, select PDB from the Choose database pull down menu. Scroll down and select Escherichia coli[orgn] from the pull down menu to the right of Limit by Entrez query. This will limit our search to only PDB records from *E. coli*. Finally, click the BLAST button to begin your search.
3. Click the Format button to retrieve your results. If they are not ready, close the window, and click Format again after a few moments.

4. On your results page, scroll down until you see the Alignment section beneath the graphic summary. Make a note of the PDB code of the *E. coli* homolog. Next, click on the red S to the right of the E-value.
5. The *T. thermophilus* protein is represented by the top bar, with the conserved FAD and photolyase domains shown beneath this. Note that the *E. coli* sequence produces two separate hits to the *T. thermophilus* sequence, one to each of the two conserved domains. Which match is the better one?
6. Click on the red bar of a hit to the FAD domain to view the alignment. Click View 3D structure to see the *E. coli* structure colored by sequence conservation with the *T. thermophilus* sequence. Residues that are identical between the two proteins will be colored red.
7. Find the FAD co-factor in the *E. coli* structure (it will be the one in the center of the protein). Following steps 1-3 in the previous section, highlight the residues within 3 angstroms of the FAD in the *E. coli* structure. The corresponding *T. thermophilus* residues will be beneath the highlighted residues in the sequence window. Point the mouse over the *T. thermophilus* residues (lower row labeled query), and find their number in the lower left corner of the window. Consult your list of residues contacting FAD in *T. thermophilus*. What residues are conserved in both proteins? What residues also contact FAD in both proteins? Are all of the residues that contact FAD in *T. thermophilus* conserved in *E. coli*? Quit Cn3D when you are finished.

Viewing a Structural Alignment of Two Photolyases

We will now build a structural alignment between the DNA photolyases from *T. thermophilus* and *E. coli*, and then attempt to discover structural differences that may lead to thermal stability.

1. Return to the NCBI home page, and click Structure on the top tool bar. Enter 1IQR in the search box to retrieve the *T. thermophilus* record. Click on the accession to load the structure summary page.
2. Click on the gray bar labeled Chain A to load the VAST structure neighbors of this protein. 1DNP, the *E. coli* protein, should be the first neighbor listed. If it isn't, either find it in the list or enter 1DNP in the box to the right of the Find button and click Find. Check the box to the left of 1DNP A (the entire chain), and click View 3D Structure to load the alignment into Cn3D.
3. Take a moment to explore the alignment, especially noting the nearly identical position of the FAD co-factor in the two structures.
4. From the menus, choose Style / Coloring Shortcuts / Secondary Structure. Now the helices are green, strands are tan, and loop/coils are blue. In the sequence window, carefully compare the two sequences, taking note of gaps (indicated by ~ marks). Which sequence has more gaps (and thus is shorter), and in what kind of secondary structure element do they occur most frequently? You can select residues from both rows in the sequence window by dragging a box over them, and this may help you to see the differences between the structures.

5. Proline rich sequences often form relatively rigid polyproline helices in proteins. In the sequence window menus, choose View / Find pattern to search for the proline repeat PPP. After clicking OK, you will need to scroll across the sequence to find any matches. Do you find a match? In which sequence is it? Are there other prolines nearby? View these residues in the structure. What might be happening here?
6. Using the evidence gathered in steps 5 and 6, form a hypothesis about how the *T. thermophilus* structure may be more stable at high temperature than the *E. coli* protein.

Identifying Bacteria in Clinical Samples using BLAST and COGs

A potentially infectious patient has made his way into the Emergency room complaining of a chronic, painful cough, fever, fatigue, night sweats and just this morning was coughing up blood.
Sputum samples were obtained and found to contain significant amounts of white blood cells and some acid-fast bacillus. Culturing of bacterial colonies is currently under way, and a sample has been sent to you at the Molecular Diagnostic Laboratory for assessment. He has been quarantined pending full diagnosis and initiation of treatment protocol.

PCR Identification of Bacterial Species

1. Using a 96-well plate of PCR primer sets reactions designed to screen bacterial cultures, your technician ran some reactions (with all of the appropriate controls). An amplified product was identified in the well with Forward primer: 5'-GTTCGGGGAGATGGAGTGCT-3' and Reverse primer: 5'-CGTTGCGGGACAGATTGATT -3'.

   Run a BLASTn on these PCR primers with a spacer (10 N's or so will do it) to identify which of the organisms in GenBank could give a band in this reaction tube.

   For example:

   ```
   GTTCGGGGAGATGGAGTGCTNNNNNNNNNNCGTTGCGGGACAGATTGATT
   ```

   ftp://ftp.ncbi.nih.gov/pub/cooper/ASMMay2003/primer.txt

   Under the BLAST button, you can "Limit by Entrez Query" to search records from a certain taxon or organism. Select Bacteria[ORGN] from the pull down menu.

   - Can you identify the species of bacteria?
   - Which gene do you think was amplified?

2. Perform a search with this gene and the genus of the infectious agent in the Database "Popset". Can you find anything of use for future studies of differentiation of the various species within this genus?

For example: "rpoB AND mycobacterium[ORGN]" retrieves:

[30145547] "Novel Polymorphic Region of the rpoB Gene Containing Mycobacterium Species-Specific Sequences and Its Use in Identification of Mycobacteria." (Lee,H. et al., J. Clin. Microbiol. 41 (41), 2213-2218 (2003))

- Can you think of ways to use this data to develop a protocol to identify particular species?

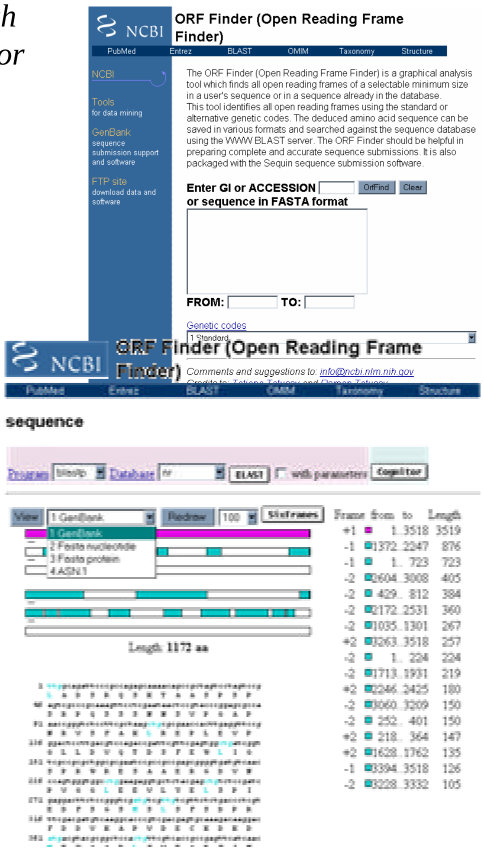
Identification of Resistance Mutations

3. Your inventive and conscientious technician sequenced the target gene from the bacterial sample. Use ORF finder (Open Reading Frame Finder) to determine the coding sequence. *ORF Finder is a graphical analysis tool which finds all open reading frames in a sequence using the standard or alternative genetic codes.*

Copy the following sequence into the text box and click on ORFfind.

`ftp://ftp.ncbi.nih.gov/pub/cooper/ASMMay2003/patient.txt`

This will take you to a graphic which will show you in shaded blocks and a table of the predicted open reading frames in all six reading frames (+1,+2,+3,-1,-2,-3). Click on the largest ORF. It will be highlighted in pink and an aligned mRNA/protein sequence will appear. *ORF finder looks for traditional ATG start sites, however in bacterial translational initiation there are occasionally alternatives to the ATG start sequence. This is the case with this particular sequence, where a TTG (usually encoding a Leucine) is the start codon.* To get the full translation, click on "Alternative Initiation Codons", "Accept" and then click on the longest ORF again. Here's your full translation. By clicking on the "View" pull-down menu, you can choose your display format of GenBank, Nucleotide or Protein FASTAs, or ASN.1.



ORF Finder (Open Reading Frame Finder)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION [ ] OrfFind Clear
or sequence in FASTA format

FROM: [ ] TO: [ ]

Genetic codes
1 Standard

Comments and suggestions to: info@ncbi.nlm.nih.gov

ORF Finder (Open Reading Frame Finder)

sequence

Frame	from	to	Length
+1	1	3518	3519
-1	1372	2247	876
-1	1	723	723
-2	2604	3008	405
-2	429	812	384
-2	2172	2531	360
-2	1035	1301	267
+2	3263	3518	257
-2	1	224	224
-2	1713	1931	219
+2	2246	2425	180
-2	3060	3209	150
-2	252	401	150
+2	218	364	147
+2	1628	1762	135
-1	3394	3518	126
-2	3228	3332	105

Next, click on "COGNITOR" to figure out what the gene and protein sequences are.

"COG" stands for ***C****luster of* ***O****rthologous* ***G****roups of proteins. These are clusters of related protein sequences which assumed to have evolved from an ancestral protein, and are therefore either orthologs or paralogs. COGnitor is a software tool that allows for a comparision of a protein sequence with the COGs database to identify the COG, if any, to which the protein belongs. Known functions (and two- or three-*

dimensional structures) of one COG member can be directly attributed to the other members of the COG. Caution must be used here, however, since some COGs contain paralogs whose function may not precisely correspond to that of the known protein.

- Based on the Heading title of the COG, what do you think is the function of this gene/protein?

From the COGNITOR page, click on “5692” (a BLAST score) next to the Rv0667 to get an alignment for the sample’s sequence to the RefSeq record. Your query sequence is on top and the reference sequence is below.

Rifampicin is a commonly used first line of defense for many antibiotic-resistant infections due to the absolute requirement of the target molecule for cell survival. This compound binds to the bacterial RNA polymerase beta and inhibits its role in transcription. The affinity of eukaryotic RNA polymerase II for rifampicin is much lower than that for this bacterial equivalent. Bacteria become resistant to this antibiotic by mutating their RNA polymerase, in particular at nucleotide positions A1344 and T1348 or C1349 which correspond to amino acid residues H445 and S450.

In your sequence, look for the key mutations which would indicate resistance to this antibiotic.

- Does this gene have the resistance mutations? What are they? (You may want to write them down to refer to them later.)

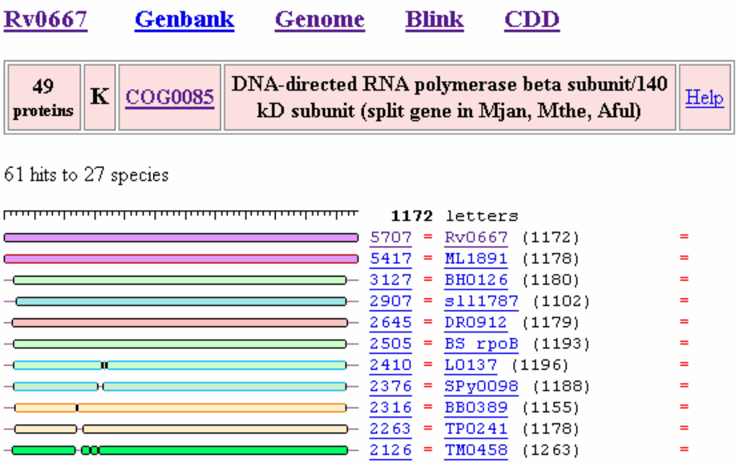
- What it your diagnosis and suggestion of treatment for the patient?

  Resistance to Rifampicin is relatively rare (<3 mutants/million organisms) and indicates a “Multidrug Resistant Strain” of the organism. In this case it is important to try high-dose combination chemotherapy to attack this disease.

Additional Exploration

4. If you have time you can do some further exploration of this enzyme and the look at the consequence of the mutations on the ability of the antibiotic to function.

From the COGs page, scroll up to the top of the page to click on the “Rv0667” to learn more about this protein.



Rv0667 Genbank Genome Blink CDD

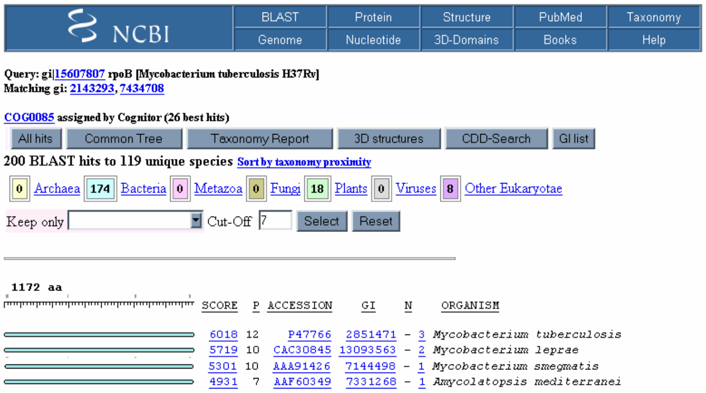
49 proteins	K	COG0085	DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful)	Help

61 hits to 27 species

```
1172 letters
5707 = Rv0667 (1172)      =
5417 = ML1891 (1178)      =
3127 = BH0126 (1180)      =
2907 = sll1787 (1102)     =
2645 = DR0912 (1179)      =
2505 = BS rpoB (1193)     =
2410 = L0137 (1196)       =
2376 = SPy0098 (1188)     =
2316 = BB0389 (1155)      =
2263 = TP0241 (1178)      =
2126 = TM0458 (1263)      =
```

This will take you to a page where you can get the FASTA formatted reference sequence ("Rv0667"), the GenBank formatted record, the graphical format of the genomic region and two other useful tools.

Clicking on "Blink" will take you to a pre-computed BLASTp page which lists the "Best hit" for this sequence in each organism in the database.



Query: gi|15607807 rpoB [Mycobacterium tuberculosis H37Rv]
Matching gi: 2143293, 7434708

COG0085 assigned by Cognitor (26 best hits)

All hits | Common Tree | Taxonomy Report | 3D structures | CDD-Search | GI list

200 BLAST hits to 119 unique species **Sort by taxonomy proximity**

0	Archaea	174	Bacteria	0	Metazoa	0	Fungi	18	Plants	0	Viruses	8	Other Eukaryotae

Keep only ▾ Cut-Off 7 Select Reset

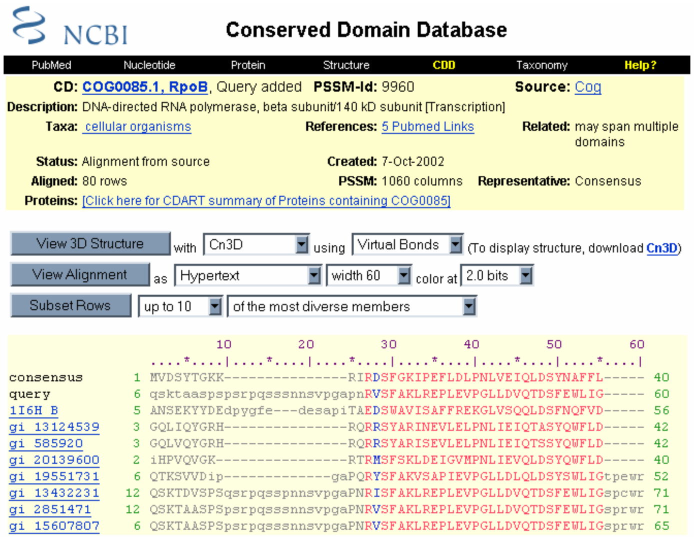
```
1172  aa
        SCORE  P ACCESSION   GI      N   ORGANISM

         6018 12    P47766  2851471 - 3 Mycobacterium tuberculosis
         5719 10  CAC30845 13093563 - 2 Mycobacterium leprae
         5301 10  AAA91426  7144498 - 1 Mycobacterium smegmatis
         4931  7  AAF60349  7331268 - 1 Amycolatopsis mediterranei
```

Clicking on "CDD" will take you to a page describing the functional domain of this protein group. *Alignments of various sequences in GenBank are shown and can be manipulated. Since this was accessed from the Rv0667 COG page, the query sequence listed here is the wild-type version.*

Search for the key residues that are mutated to confer the Rifampicin resistance.

- Are the key residues (H445 and S450) located within a conserved region based on the alignment?



Conserved Domain Database

CD: **COG0085.1, RpoB**, Query added **PSSM-Id:** 9960 **Source:** Cog
Description: DNA-directed RNA polymerase, beta subunit/140 kD subunit [Transcription]
Taxa: cellular organisms **References:** 5 Pubmed Links **Related:** may span multiple domains
Status: Alignment from source **Created:** 7-Oct-2002
Aligned: 80 rows **PSSM:** 1060 columns **Representative:** Consensus
Proteins: [Click here for CDART summary of Proteins containing COG0085]

View 3D Structure with Cn3D ▾ using Virtual Bonds ▾ (To display structure, download Cn3D)
View Alignment as Hypertext ▾ width 60 ▾ color at 2.0 bits ▾
Subset Rows up to 10 ▾ of the most diverse members ▾

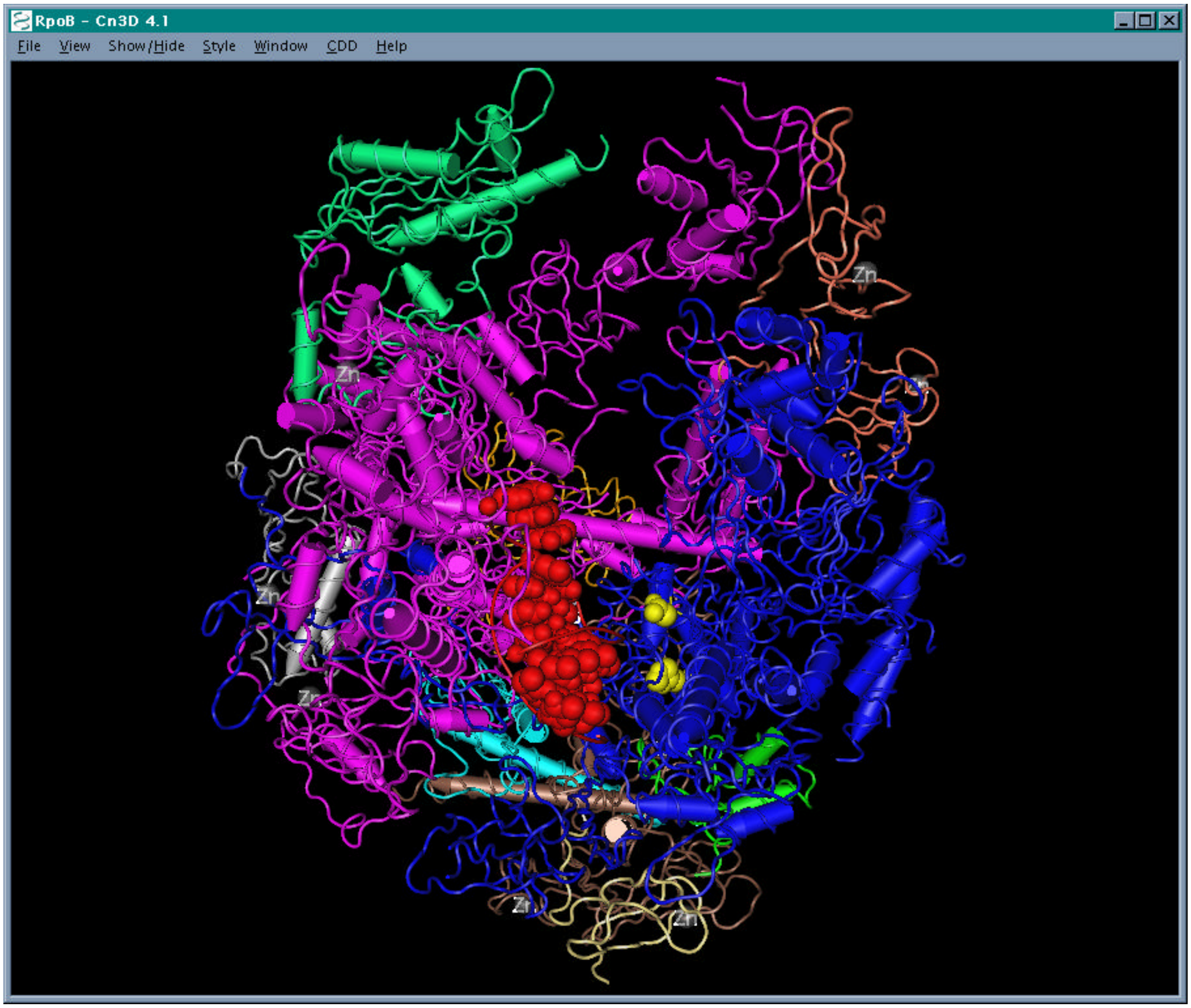
```
                     10        20        30        40        50        60
                ....*....|....*....|....*....|....*....|....*....|....*....|
consensus     1 MVDSYTGKK---------------RIRDSFGKIPEFLDLPNLVEIQLDSYNAFFL----- 40
query         6 qsktaaspspsrpqsssnnsvpgapnRVSFAKLREPLEVPGLLDVQTDSFEWLIG----- 60
1I6H B        5 ANSEKYYDEdpygfe---desapiTAEDSWAVISAFFREKGLVSQQLDSFNQFVD----- 56
gi 13124539   3 GQLIQYGRH---------------RQRRSYARINEVLELPNLIEIQTASYQWFLD----- 42
gi 585920     3 GQLVQYGRH---------------RQRRSYARISEVLELPNLIEIQTSSYQWFLD----- 42
gi 20139600   2 iHPVQVGK----------------RTRMSFSKLDEIGVMPNLIEVQLDSYQWFLD----- 40
gi 19551731   6 QTKSVVDip-------------gaPQRYSFAKVSAPIEVPGLLDLQLDSYSWLIGtpewr 52
gi 13432231  12 QSKTDVSPSqsrpqsspnnsvpgaPNRISFAKLREPLEVPGLLDVQTDSFEWLIGspcwr 71
gi 2851471   12 QSKTAASPSpsrpqsssnnsvpgaPNRVSFAKLREPLEVPGLLDVQTDSFEWLIGsprwr 71
gi 15607807   6 QSKTAASPSpsrpqsssnnsvpgaPNRVSFAKLREPLEVPGLLDVQTDSFEWLIGsprwr 65
```

Another feature of the CDD page is that if a structure that has been identified as containing this particular domain, you can access the record to see what general structure of the domain.

In this case there is a structure from the *Saccharomyces cerevisiae* RNA polymerase II which is complexed with the entire Transcriptional Elongation Complex. Change "Virtual Bonds" to "All Atoms", and then click on "View 3D Structure" to see the RNA polymerase II domain with the alignment (including your Rv0667 query sequence) in the Sequence/Alignment viewer.

You can search for the mutated residues by scrolling along the sequence and placing your cursor over the query sequence's letters. In the bottom of the viewer you'll see the position of that particular residue. Find H445 and S450. Click on them (holding the control key will allow you to select both) and they will turn yellow in both the Sequence/Alignment viewer and in the Molecule Window.

To see how they fit within the entire Transcriptional Elongation Complex, click on "Show/Hide" then "Show everything". You may need to zoom out (type "x", to zoom in hit "z"). Color the molecules by clicking on "Style" then "Coloring Shortcuts" then "Molecule". *The blue structure is the RNA polymerase II and the highlighted mutated residues should still be seen in yellow. These residues are where the Rifampicin binds to the bacterial enzyme. You should be able to see the DNA/RNA duplex where it needs to bind in the central core of the enzyme (shown here in red spacefill).*



- Can you suggest why the antibiotic works so well?